**EXPLORATORY DATA ANALYSIS**

**Outcome Variables**

There are three response variables representing count of all renters, count of registered renters, and count of casual renters. Registered and casual renters are two separate customer bases. From the below summary statistics, it’s not hard to see that the overall demand among registered renters are far higher than that of the casual renters. Therefore, we examine the distribution separately for the two groups.

Descriptive Statistics of Registered, Casual, and Total Rentals

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| variable | mean | median | sd | min | max |
| casual | 22.56 | 3 | 43.21 | 0 | 367 |
| registered | 97.44 | 23 | 141.25 | 0 | 886 |
| count | 120 | 28 | 170.71 | 0 | 977 |

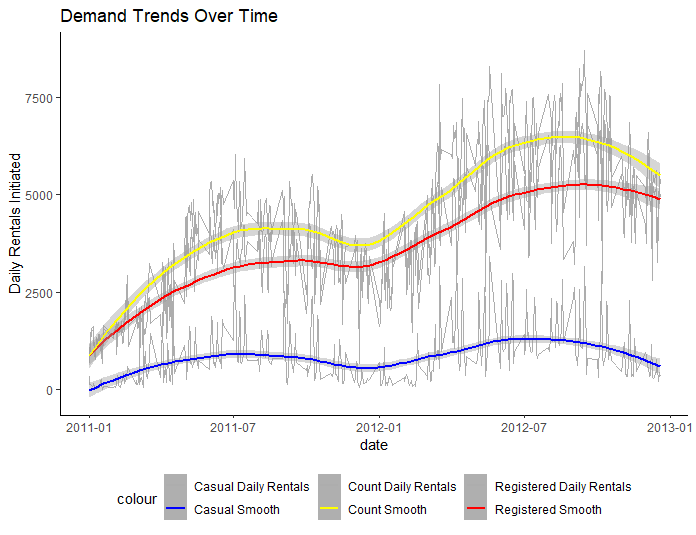
**Predictors**

The observations in the train and test sets are combined to get an overall distribution of the predictors in the dataset. Predictor variables of holiday, season, weather and workingday are converted to factor variables in r. Based on the distribution, we see that majority of the observations are non-holidays /workdays, and have weather as “Clear, Few clouds, Partly cloudy, Partly cloudy”. Notably, there are no “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog” in the dataset. Season has four categories of almost equal distribution. Variables temp, atemp, humidity look approximately normally distributed, thought temp displays a slightly bimodal distribution. Windspeed is right skewed.

|  |  |
| --- | --- |
| Distribution of Numeric Predictors | Distribution of Categorical Predictors |
|  |  |

**Time Trend**

Overall, rental demands gradually increase from 2011 to 2012. Registered rentals account for the majority of the increasing trend while casual rentals stay relatively stable across the two years. There’s also a distinct seasonal trend of increasing bike rides going from winter to summer.

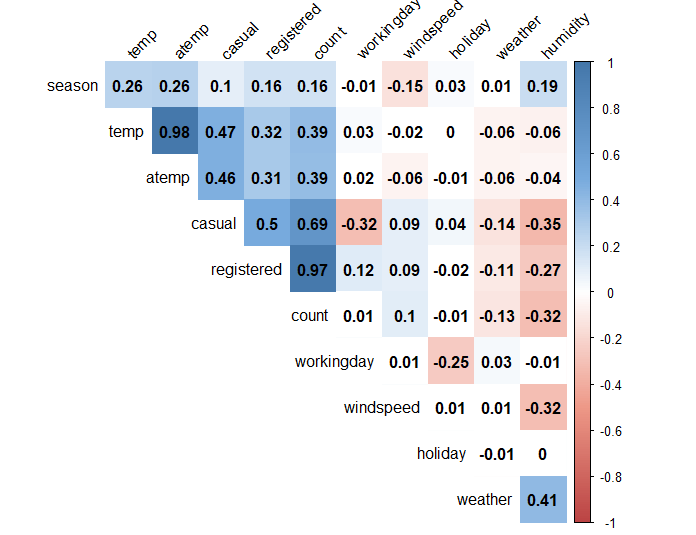


Year, month, date of the week, and hour are created from the datetime variable to further examine potential time trend in rental demands. There’s a higher bike demand in year 2012 as compared to 2011 for both registered and casual. The demand is also higher in quarter 2 and 3 compared to quarter 1 and 4. For casual riders, the mean and variance in demand on the weekends are higher than on weekdays, while the exact opposite is found in the mean and variance for registered rentals. The hourly plot shows that demands are high in 7-9 and 17-19 hours for registered renters, around average in 10-16 hours, and low in 0-6 and 20-24 hours. Demands for casual rentals do not exhibit distinct peak hours, instead it gradually increases from 6-17 hours and decreases in 18-24 hours and the ride counts are relatively high throughout the mid-day hours. From the time trends, it’s not hard to infer that registered customers typically use the Capital Bikeshare system to transport to work/school while casual riders are mostly tourists who utilize the Bikeshare system to navigate the city.

|  |  |
| --- | --- |
|  |  |
|  |  |

**Correlations**

The correlation plot shows that temp is positively correlated to the ride counts for both registered and casual riders. On the other hand, humidity and weather are negatively correlated to the outcome variables. Windspeed is also positively correlated to ride counts though the correlation is a lot weaker compared to temp, humidity, or weather. As expected, temp and atemp are highly positively correlated. Workingday is positively correlated to registered rentals while negatively correlated to casual rentals. Conversely, holiday is positively correlated to casual rentals and negatively correlated to registered rentals. This matches our previous inference that the two groups capture professional and tourist customer bases, respectively.



**Transforming Response Variables**

From the boxplot of registered and casual variables, it is clear that the distribution of the counts highly skewed for both registered and casual rentals. However, in evaluating transformations of this variable, it is necessary to consider how it will affect the relationships with all the other variables since registered and casual will be the response variable in the model. For this reason, the distributions of all numeric variables are compared after the transformation has been applied to the response variable. In a useful transformation of the response, we expect to see the residual distribution of most variables improve (become more mound-shaped). Log, Square-root, and inverse transforms are evaluated. It is pretty clear from the residual distributions that a square root of the response is helpful.

Transformation of Registered Rentals

|  |  |
| --- | --- |
| No transformation | Log transformation |
|  |  |
| Square root transformation | Inverse transformation |
|  |  |

Transformation of Casual Rentals

|  |  |
| --- | --- |
| No transformation | Log transformation |
|  |  |
| Square root transformation | Inverse transformation |
|  |  |